Efficient Range Distribution Query in Large-scale Scientific Data

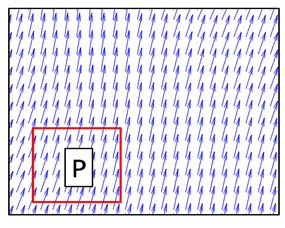
Abon Chaudhuri, Teng-Yok Lee, Han-Wei Shen
The Ohio State University

Tom Peterka
Argonne National Laboratory

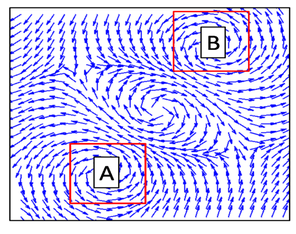
Range Distribution Query

- Query-based visualization is useful for exploring large-scale data
- Distributions can substitute for raw data to answer statistical queries
- ***Range Distribution Query:*** Given an axis-aligned region defined by a spatial range, return the distribution within that region
  - Enables interactive exploration
  - Benefits visualization/analysis applications that need histograms at various levels of detail
- Response to range distribution query is a function of query size which makes it slow and space-consuming from large-scale data
- ***Main goal:*** to answer such queries in constant time regardless of size

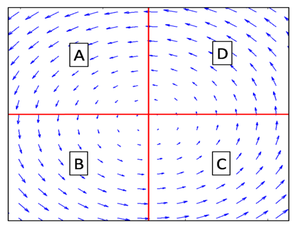
Key Intuition: Similarity and Redundancy within Data



Sub-region P is similar to the entire region



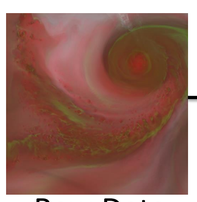
Sub-region A is similar to sub-region B



Sub-regions A, B, C and D are transformable to each other

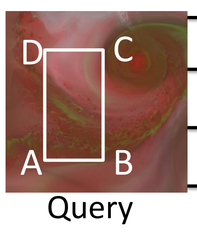
- We observe that scientific data contains redundancy of various types
- Distributions (represented as histograms) computed from such data are also similar to each other, or at least transformable from one to another

Proposed Framework



Raw Data → Integral Distribution Volume (IDV) → Hierarchy of re-usable sub-range distributions → ***Similarity based indexing*** → Error/Residual, Codebook → ***LZ/BZ2/RLE/ Wavelet Compression*** → Compressed Error/Residual

Pre-preprocessing phase



Query → List of blocks to reconstruct H(A), H(B), H(C), H(D) → Compressed Error/Residual, Codebook → H(A), H(B), H(C), H(D) → +/- → h(ABCD)

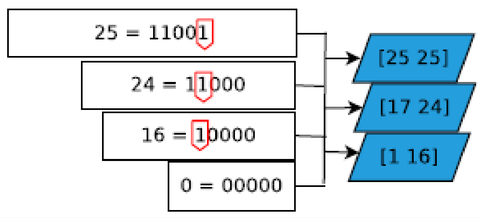
Query processing

- Our framework leverages the use of a summed area table (SAT) like data structure called ***Integral Distribution Volume***
- We exploit the similarity among different regions in the data and their respective distributions (histograms)
- This allows ***similarity-based indexing*** of the distributions by template distributions obtained from the data itself, leading to more efficient compression of the error / residual
- Our method increases compressibility of the integral histograms and can be used with any off-the-shelf compression technique

Integral Distribution Volume (IDV)

- Given a 3D field, IDV stores at each grid point the prefix histogram or the ***integral histogram*** – the histogram of the region spanning from origin to that point [1]
- The ***benefit of IDV*** is that the histogram of any query region can be computed in a constant time
- The ***problem of IDV*** is its huge storage cost - it consumes storage equal to *k* times the raw data size (when *k*-bin histograms are used)

Decomposition of Ranges

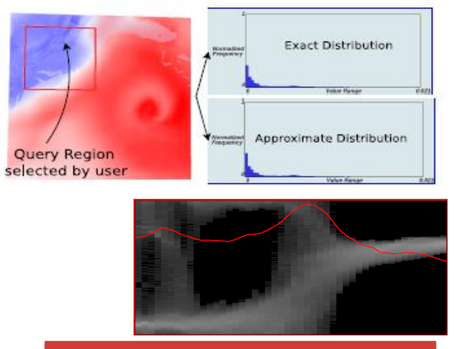


- Any range [1,P] can be decomposed into a small number of (less than $\log_2 P$) sub-ranges with power-of-two length using a fast bitwise operation based algorithm [2]
- The principle applies to 2D or 3D sub-ranges

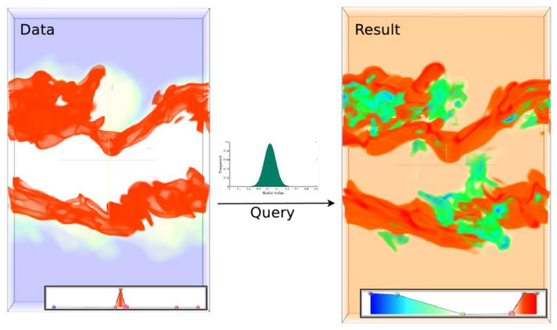
Similarity-based Indexing

- **Input:** set of sub-range distributions B and a much smaller set of template distributions T (B>>T)
- The indexing algorithm maps each sub-range distribution h_B with a template distribution h_T which best approximates h_B under some optimal transformation (shift and/or reflection in our case)
- The template set T along with each mapping is stored in a data structure called ***codebook***, which is later accessed to reconstruct the sub-range distributions, and hence the query distributions

Applications



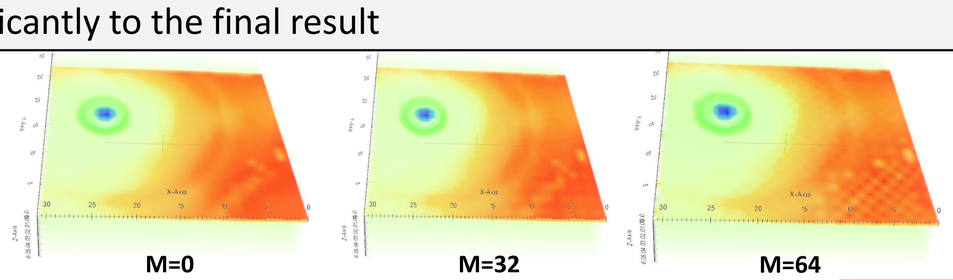
Interactive visualization



Distribution-based search

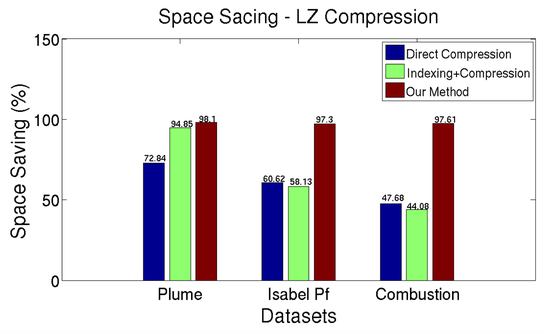
Summary statistics fields: Mean, variance, entropy etc. can be computed at different levels of detail by retrieving block distributions

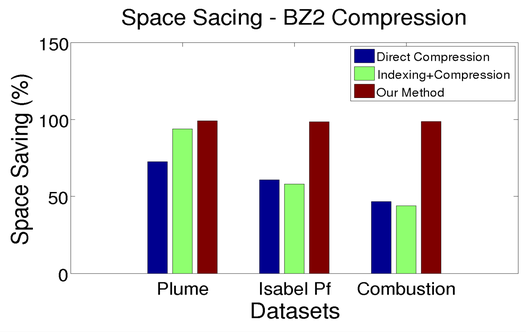
Controllable query workload: The query workload can be controlled by selectively discarding sub-range distributions which do not contribute significantly to the final result



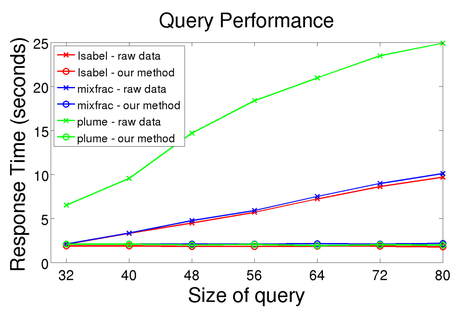
M=0 **M=32** **M=64**

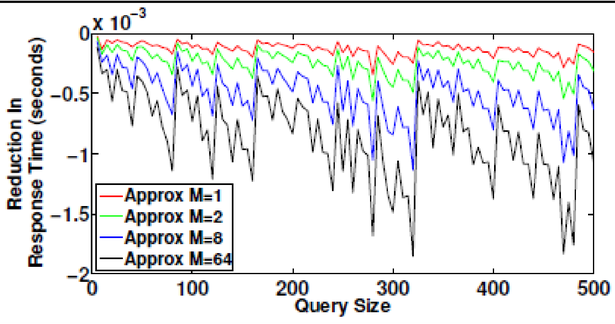
Quantitative Analysis





- The space-saving is measured compared to the projected size of IDV





- Our method is faster than raw data access for large queries on blocked data
- We can achieve progressively faster query response by selectively discarding blocks with negligible contribution to the query region histogram

Conclusions and Future Work

- This poster presents a technique to support arbitrary range distribution query on volumetric data
- Indexing and query processing are easy to parallelize
- Plan to use this technique for specific analysis tasks such as distribution-driven streamline computation

Acknowledgments

- Isabel is benchmark for IEEE Vis Contest 2004
- NCAR scientists for Plume dataset
- Sandia National Lab for Combustion dataset
- NSF for grants IIS-1017635 and IIS-1065025, US DoE for DOE-SC0005036, Battelle Contract No. 137365, and SciDAC for grant DE-FC02-06ER25779.
- Program manager Lucy Nowell

References

- [1] F. Porikli. Integral histogram: A Fast Way to Extract Histograms in Cartesian Spaces. IEEE Conference on Computer Vision and Pattern Recognition, volume 1, pp. 829 – 836, 2005.
- [2] S.Martin and H.-W. Shen. Transformations for volumetric range distribution queries. IEEE Pacific Visualization Symposium, 2013